ECE 20875 Mini Project Report

Team Members:
Luke Ritchison, lritchis@purdue.edu, lritchis, path 2
Julian Knudson, jknudso@purdue.edu, jknudso, path 2

The data we are working with:
We are working with the behavior-performance.txt document, which contains data pertaining to students' viewing behavior for online lectures for a course. This data includes:

- The ID of the students
- The ID of the video the student is watching
- The amount of time spent on the video (relative to video length)
- The fraction of the video watched
- The amount of time spent paused (relative to video length)
- The number of times the student pauses the video
- The average playback rate of the video
- The number of times the video was rewind-ed
- The number of times the video was fast-forwarded
- Whether the student answered the quiz question correctly

For the respective questions, we will be using the following analyses:

For the first question, we will be performing KNN analysis on some of the given data points with respect to their peers. We will then measure the accuracy (fraction of correct predictions) and determine how well the KNN model predicts which cluster a specific point will fit into. If the accuracy is high, we will conclude that the students can be grouped/clustered well based on the respective variable, and vice versa. This is because, the better the clustering, the less often we will see foreign neighbors appear next to each other.

For the second question, we will take each of the variables, and plot them against test scores (both being averaged with respect to student ID's). From these plots, we will use polynomial regression to create models to fit these scatterplots of data. We will use degrees of 1, 2, 3, 4, and 5 for the polynomials. To avoid overfitting, we will avoid using degrees greater than 5.

Once these models are made, we will find the r-squared values to determine goodness-of-fit for each model. The better the fit, the better we can predict a student's performance on a question, and vice versa. This is because a higher r-squared for a specific variable means more of the variation in the scores can be attributed to that variable.

If multiple models have strong (> 0.5) r-squared values for the same dataset, we will use training and testing data to determine which model to use. We cannot use the r-squared values mentioned above to choose a model, because higher degree models will always have a better training r-squared value, due to overfitting. But, if we make 90% of the data into training data, and 10% into testing data, we will be able to prevent our decision from being corrupted by overfitting.

For the third question, we are going to follow the same process as problem 2, but using the average data for each video, as opposed to the averages per student. We will do this because video behavior among all students is the best data to use to predict a new student's performance on that video. Again, we will make models of degrees 1, 2, 3, 4, and 5; we will calculate the r-squared values; and we will use training and testing data to determine which model is best, if multiple models for a single variable have a strong (> 0.5) r-squared value.

Analysis results for question 1:

For question 1, the data was filtered out based on the appearance of student IDs within the dataset. If a student ID appeared 5 times or more, it means that student interacted with 5 or more videos, and was deemed appropriate to keep in the dataset for this analysis.

Once that filtering was completed, each student then had each of their video parameters used to create an average. That is, across all parameters, each datapoint now represented a given student's average parameters. With this dataset, it would now be possible to draw conclusions about students' habits on a broader scale, rather than on an individual video basis.

From there, to perform a K-nearest neighbors analysis, it was required to have an objective metric by which to form the data labels. Score was chosen to be that metric, since it was not a variable involved in the analysis and it was deemed the most important metric to be able to describe or predict. Since score (now averaged across at least 5 videos) ranges from 0 to 1, the categories were split into fourths of that range, being (0, .25], (.25, .5], (.5,.75], (.75,1]. Each data point was assigned its appropriate label.

The data was then split into 90% training data and 10% test data. The sk.neighbors module from sklearn was used to perform the fitting of the training data and evaluate the accuracy of the models. The analysis was performed for 1 to 5 nearest neighbors, and across the variables *fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs,* and *numFFs*.

None of the models held up well. Recall that accuracy is measured from the confusion matrix, which is constructed as a list of predicted categories vs objective categories. If a predicted category is anything but a match for the objective category, that will be indicative of a less accurate model. The highest accuracy was achieved for 5 nearest neighbors, being .3072. It was followed very closely by 1 nearest neighbor. The other results can be seen in the table below.

Table 1: K-Nearest Neighbor Accuracy Results

Nearest Neighbors	Accuracy
1	0.3006535947712418
2	0.23529411764705882
3	0.27450980392156865
4	0.29411764705882354
5	0.30718954248366015

As shown, there was never a high degree of accuracy achieved. We have concluded on this basis that, at least on the basis of average score, the other variables cannot be naturally grouped well. This indicates that these variables are not good predictors of a student's score.

Analysis results for question 2:

The loading of the .csv file containing the data is found in main.py. All other code for problem 2 is found in problem2.py.

In the file problem2.py, we first filtered the data. The question specifies that we need average data for the viewing categories for students who participated in at least half of the quizzes. Since there are 93 quizzes, we rounded up 93/2 to get students who appeared at least 47 times in the dataset. There were 94 such students.

Then, we used the feature matrix to come to a least-squares model at degrees 1 through 5 for the average score plotted against each of the other 7 values (averaged) for the 94 students. Following this step, we plotted the experimental data against the respective models, and found that three of the variables, (fracSpent, fracPaused, and numRWs), had visible outliers. All the plots we got from this step are found in Figures 1-7 in Appendix A.

After this, we retroactively removed these outliers from their respective datasets, and came to the models that we would use to calculate r-squared. The plots of the models can be found in Figures 8-14 in Appendix B, and the model coefficients and r-squared values can be found in Figures 15-21 in Appendix C.

Since none of the 35 models of the 7 variables had strong (> 0.5) r-squared values, we have no need to use training and testing data to determine which model to use. In fact, none of the models even surpassed an r-squared value of 0.15, and 23 of the 35 models failed to surpass 0.05.

Based on the low r-squared values across the board, we conclude that none of the models generated for these seven variables in relation to score explain more than 15% of the variability in the data; therefore, we conclude student's video-watching behavior can not be used to predict a student's performance.

Analysis results for question 3:

Everything in the analysis for question 2 is the same here, with the following exceptions:

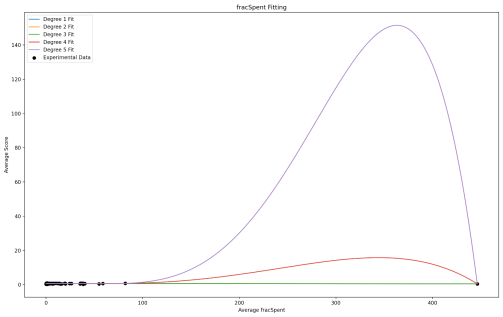
There were 92 data points, because there were 93 quizzes but quiz #29 never appeared, therefore 93 - 1 = 92. There were no outliers to remove, so the plots are found in Figures 22-28 in Appendix D, and the model coefficients and r-squared values are found in Figures 29-25 in Appendix E.

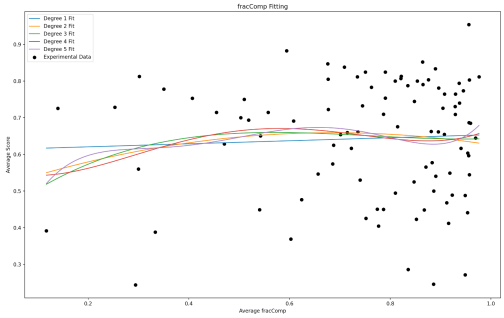
Since none of the 35 models of the 7 variables had strong (> 0.5) r-squared values, we have no need to use training and testing data to determine which model to use. In fact, none of the models even surpassed an r-squared value of 0.15, and 23 of the 35 models failed to surpass 0.05. (Quite a coincidence that both problem 2 and problem 3 have 23 models under 0.05. Just an interesting observation.)

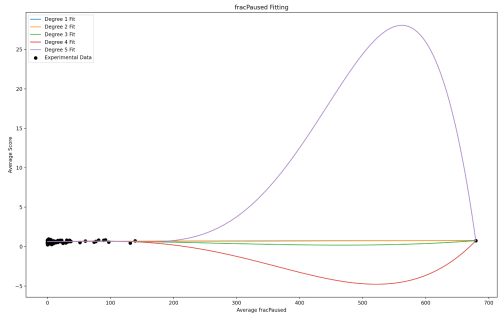
Based on the low r-squared values across the board, we conclude that none of the models generated for these seven variables in relation to score explain more than 15% of the variability in the data; therefore, we conclude that video-watching behavior can not be used to predict student's quiz score on a specific video.

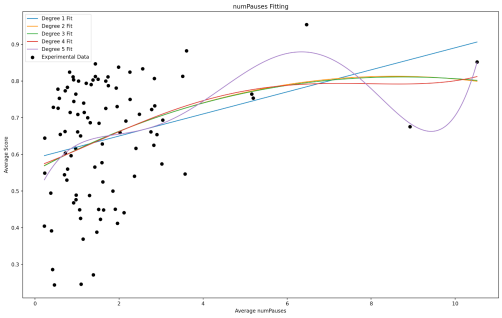
Appendix A

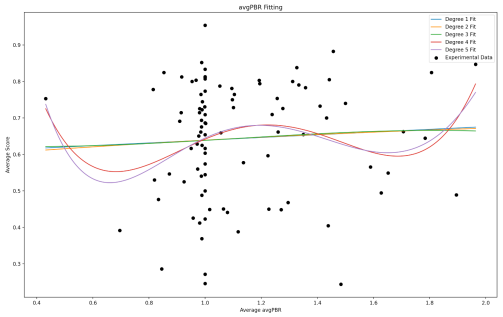
Figures 1-7: Problem 2 Model Plots w/ Outliers

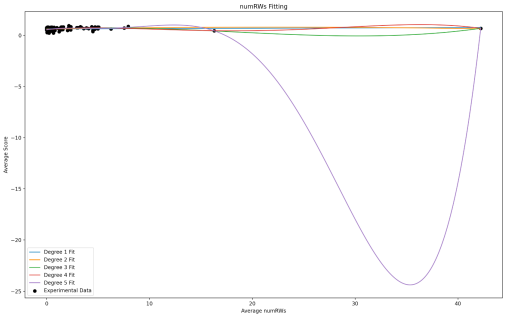


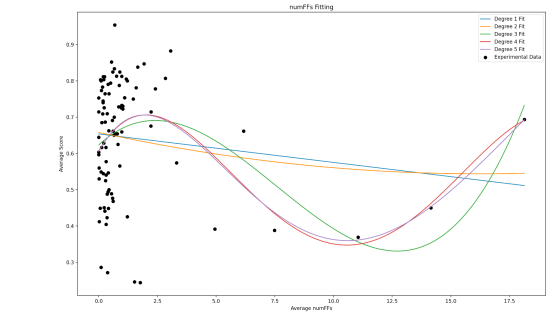






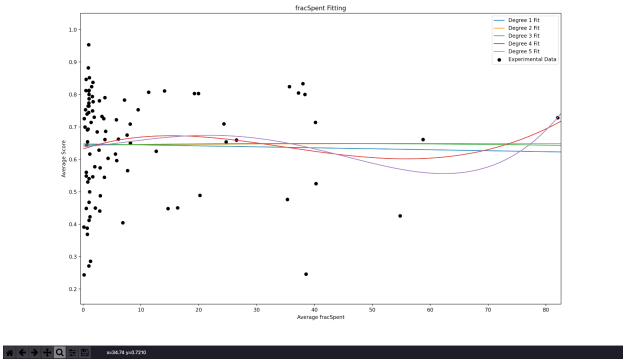


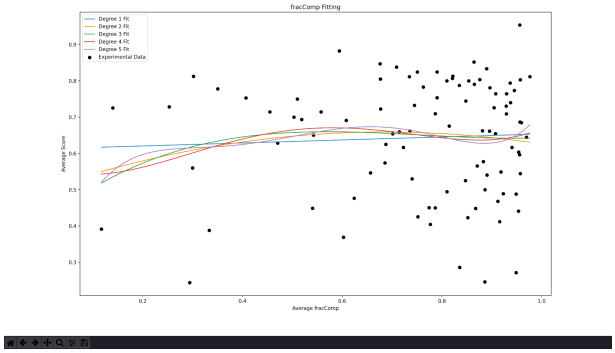


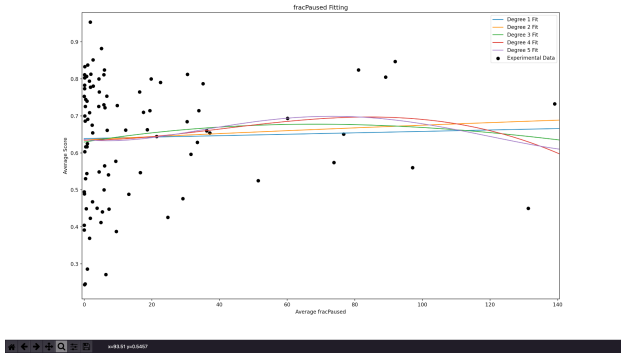


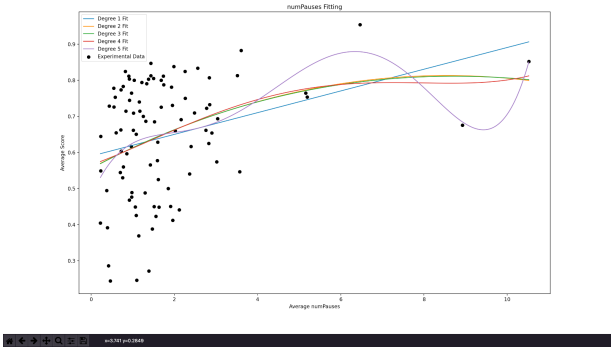
Appendix B

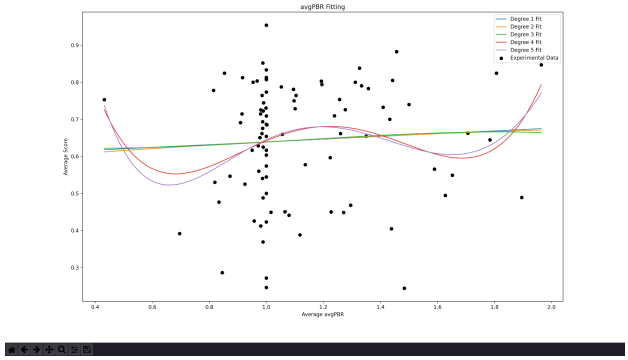
Figures 8-14: Problem 2 Model Plots w/o Outliers

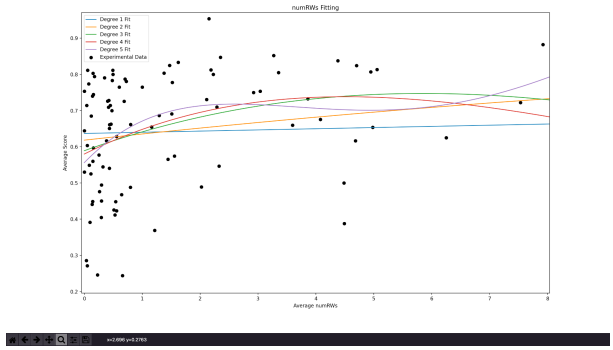


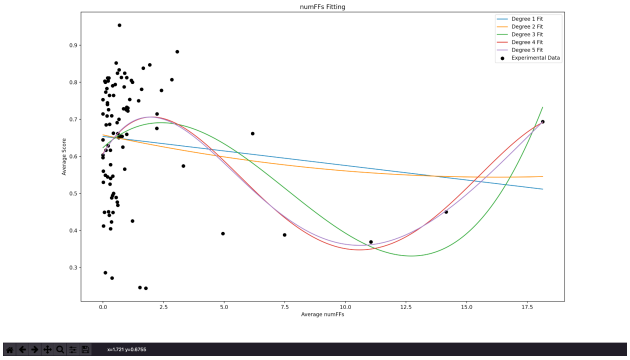












Appendix C

Figures 15-21: Problem 2 Model Coefficients and R-Squared Values

```
The amount of time spent on the video (relative to video length)
8.94794303115466e-05 X + 0.6441998304645599
R-Squared Value: 7.286156752173145e-05

-7.733146359532717e-07 X^2 + 0.0001318630091414314 X + 0.6440484258543482
R-Squared Value: 7.572904538599712e-05

2.155724634314188e-06 X^3 + -0.0002347959674675l012 X^2 + 0.005768751151462512 X + 0.6320049278155335
R-Squared Value: 0.010315857724854105

5.495623487632348e-08 X^4 + -5.822083451417185e-06 X^3 + 0.0001020528229706679 X^2 + 0.0015852665353293656 X + 0.6382999286570069
R-Squared Value: 0.012662991992301587

1.5381586307834582e-09 X^5 + -2.2568312235483374e-07 X^4 + 1.1388748014257597e-05 X^3 + -0.0003076686210358938 X^2 + 0.004708664498200737 X + 0.6347614343686722
R-Squared Value: 0.01308132099870829
```

```
The fraction of the video watched
0.041395574394463654 X + 0.6122187617768472
R-Squared Value: 0.002828768567885609

-0.3371277967746169 X^2 + 0.4626802568193693 X + 0.5002907772859317
R-Squared Value: 0.013183260400560615

0.6752666236688405 X^3 + -1.5368119719471232 X^2 + 1.094287050873478 X + 0.41025105986076343
R-Squared Value: 0.015346320628423116

3.179513049265684 X^4 + -6.5878094560796185 X^3 + 4.1396008085196385 X^2 + -0.6327387798323021 X + 0.570540063449315
R-Squared Value: 0.017716585423199982

20.47896927406358 X^5 + -55.00323302931258 X^4 + 55.00704665212592 X^3 + -25.541655724901986 X^2 + 5.646119438297795 X + 0.1307013861975601
R-Squared Value: 0.021512205739708246
```

```
The amount of time spent paused (relative to video length)
0.00036869807108821397 X + 0.6360910582713637
R-Squared Value: 0.004318147551682028

-1.0155637528884187e-05 X^2 + 0.001425399000129437 X + 0.6293832843580209
R-Squared Value: 0.009233298547382418

-1.7735455995565638e-07 X^3 + 2.1904581906395254e-05 X^2 + 0.0001585841065907659 7 X + 0.633895962920359
R-Squared Value: 0.01146861472591787

4.0521882999087216e-09 X^4 + -1.182913681569737e-06 X^3 + 9.416491903240259e-05 X^2 + -0.0012661316549529142 X + 0.636852423020484
R-Squared Value: 0.01222335416315612

3.716195048753396e-10 X^5 + -1.100343765965514e-07 X^4 + 1.0630291698124963e-05 X^3 + -0.0003830102762358665 7 X^2 + 0.004976895045901173 X + 0.6270131415638355
R-Squared Value: 0.021991027809108
```

```
The number of times the student pauses the video
0.030105462764951376 X + 0.589730363124803
R-Squared Value: 0.09163983273965748

-0.0035132707281828653 X^2 + 0.060050080876867085 X + 0.5564571194408837
R-Squared Value: 0.10601260356109876

4.233645932446893e-05 X^3 + -0.0041008160220875525 X^2 + 0.06189569159851219 X + 0.5551799094604122
R-Squared Value: 0.10602497435617408

9.308077006961757e-05 X^4 + -0.0017502572920935103 X^3 + 0.006355591130776137 X^2 + 0.041957363746549195 X + 0.5650635591576095
R-Squared Value: 0.10644355877262734

0.00023769508554740876 X^5 + -0.005786818226840967 X^4 + 0.04836725081857148 X^3 + -0.16776004802625072 X^2 + 0.2736116641690942 X + 0.4769808088774734
R-Squared Value: 0.12728088675765314
```

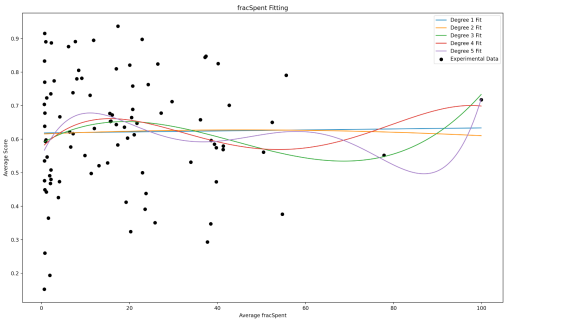
```
The average playback rate of the video
0.03719891413720511 X + 0.6016468354648434
R-Squared Value: 0.003808979113906541

-0.009510696092658158 X^2 + 0.061288352732741955 X + 0.587248202469731
R-Squared Value: 0.00384707724698318 8

-0.032226453169123995 X^3 + 0.10795590491261231 X^2 + -0.0733501674504845 X + 0.635969533333612
R-Squared Value: 0.00395498488143009

1.6618561565103454 X^4 + -7.955814602013791 X^3 + 13.442908202831918 X^2 + -9.343633521653674 X + 2.836506102461835
R-Squared Value: 0.0412588837758604

-0.9118087026159953 X^5 + 7.18170971565082 X^4 + -20.743787754057063 X^3 + 27.504875167295637 X^2 + -16.595894739881118 X + 4.209763654873949
R-Squared Value: 0.04288444379843881
```

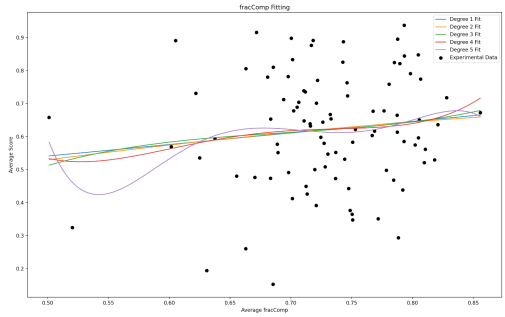
```
The number of times the video was rewind-ed
0.026630547813923235 X + 0.6059975982747612
R-Squared Value: 0.08913480389844908

-0.005492690056531406 X^2 + 0.05877449777665929 X + 0.5882051551500802
R-Squared Value: 0.10661625332789892

0.003412321871789005 X^3 + -0.04184926126540739 X^2 + 0.14999664788407893 X + 0.5581603335549227
R-Squared Value: 0.13820641783791643

0.0011257874028049212 X^4 + -0.012672508142467884 X^3 + 0.026553795645570887 X^2 + 0.061589310653104024 X + 0.5770985019119128
R-Squared Value: 0.1476348178469168

-0.0001469431626704143 2 X^5 + 0.003751755286613683 X^4 + -0.028720298009915518 X^3 + 0.06579036680839112 X^2 + 0.02854322678055743 X + 0.5824076833925842
R-Squared Value: 0.1483981231529553
```

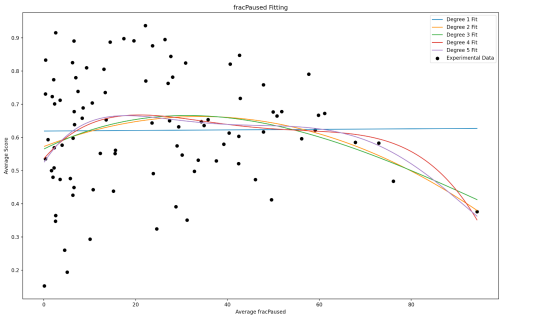
```
The number of times the video was fast-forwarded
-0.007845224676983209 X + 0.6536248415373319
R-Squared Value: 0.018058129110834154

0.0004323222571804054 X^2 + -0.014045322654284566 X + 0.657746007921552
R-Squared Value: 0.019459474571902557

0.0006550991988691141 X^3 + -0.014853421710890016 X^2 + 0.05993975073831992 X + 0.6233320209729387
R-Squared Value: 0.09924748314783916

-6.662696914076778e-05 X^4 + 0.0028213110905437365 X^3 + -0.034996080855614904 X^2 + 0.10881537263628582 X + 0.6069536935146442
R-Squared Value: 0.11492016348061262

1.926801213437728e-06 X^5 + -0.0001444767794197597 3 X^4 + 0.003861572472958577 X^3 + -0.04015242290318143 X^2 + 0.11631036896405123 X + 0.6051382139522183
R-Squared Value: 0.11509315726064084
```

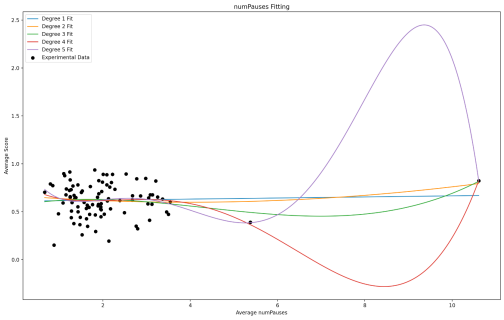
Appendix D

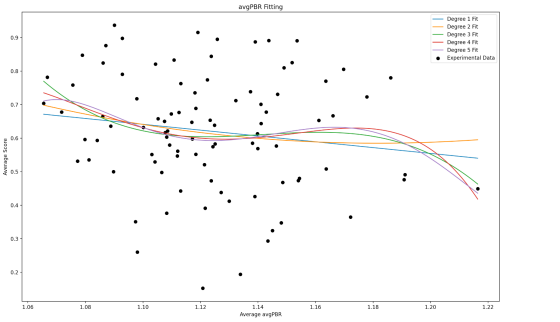
Figures 22-28: Problem 3 Model Plots

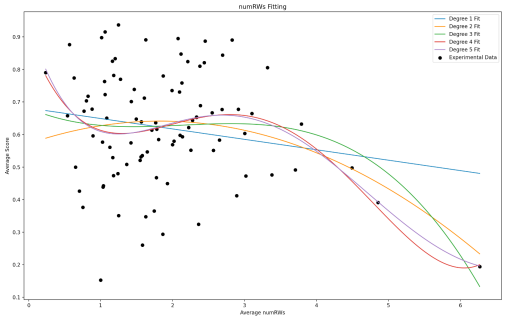


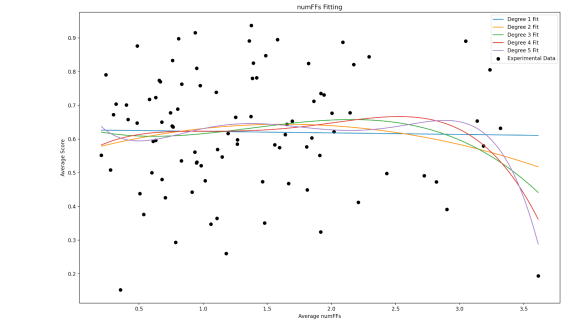












Appendix E

Figures 29-35: Problem 3 Model Coefficients and R-Squared Values

```
The amount of time spent on the video (relative to video length)
0.00015129743807337757 X + 0.6183979550855773
R-Squared Value: 0.0002645182718453709

-5.85737638644171e-06 X^2 + 0.000535545903939839 X + 0.6152930827541241
R-Squared Value: 0.0006652572124136169

1.9291821581033244e-06 X^3 + -0.0002539487872487762 X^2 + 0.007580986001490489 X + 0.5867472098827692
R-Squared Value: 0.029315395708814695

-3.929793423638931e-08 X^4 + 8.759059396398477e-06 X^3 + -0.000599983459196224 X^2 + 0.01295014603873146 X + 0.5741940042251948
R-Squared Value: 0.035580536340305224

2.7975250543593382e-09 X^5 + -6.739746743699026e-07 X^4 + 5.7722570982539835e-05 X^3 + -0.0020999851910326836 X^2 + 0.0287451916214734 X + 0.5485808604729436
R-Squared Value: 0.05052491655052882
```

```
The fraction of the video watched
0.35010279384446485 X + 0.3656769981841785
R-Squared Value: 0.016301557092774144

-0.3228100786521534 X^2 + 0.8025593254968972 X + 0.20866261520148296
R-Squared Value: 0.016457559045032655

7.205996325464565 X^3 + -15.034367934102413 X^2 + 10.689997490101174 X + -1.9751688792537725
R-Squared Value: 0.017093554346427564

128.8763801779021 X^4 + -345.26181022738905 X^3 + 343.17375888856753 X^2 + -149.49757068658846 X + 24.592420983520583
R-Squared Value: 0.018713674997266394

-3193.1066948255257 X^5 + 11161.661716858385 X^4 + -15484.943588948609 X^3 + 10652.549152721429 X^2 + -3631.4844247784545 X + 490.99763723308223
R-Squared Value: 0.025532893161767545
```

```
The amount of time spent paused (relative to video length)
8.499636489664446e-05 X + 0.6191293598821924
R-Squared Value: 0.00010805902592481154

-7.793460336914211e-05 X^2 + 0.0053121189393559465 X + 0.5724893105126581
R-Squared Value: 0.058546831739331284

4.25896865345971e-07 X^3 + -0.00013037951242953247 X^2 + 0.006894668140738221 X + 0.5648132869040379
R-Squared Value: 0.05973961113066251

-5.6008140004928824e-08 X^4 + 1.0212202740692424e-05 X^3 + -0.000656024480162761 X^2 + 0.016073404036179233 X + 0.5356440224996174
R-Squared Value: 0.07141261796640641

9.48408874539405e-10 X^5 + -2.6327768022949644e-07 X^4 + 2.5837206159355883e-05 X^3 + -0.0011353542483824512 X^2 + 0.02138305793122045 X + 0.5236754331281663
R-Squared Value: 0.07310304428349124
```

```
The number of times the student pauses the video
0.005645657706401345 X + 0.6094873846453696
R-Squared Value: 0.0015423713377593629

0.004392646294149545 X^2 + -0.03474537094510569 X + 0.6680808069740753
R-Squared Value: 0.015662813625328176

0.0024138287339346676 X^3 + -0.03158613918183903 X^2 + 0.08768772067414406 X + 0.559839604777735
R-Squared Value: 0.02521102654338847

0.002059889076378433 X^4 + -0.03598236484690444 X^3 + 0.1829478166421438 X^2 + -0.35246699138917353 X + 0.843280007100038
R-Squared Value: 0.04350019664989113

-0.0012678876135717749 X^5 + 0.02962805261477657 X^4 + -0.2380407000365698 X^3 + 0.8139667635718919 X^2 + -1.1985674025530306 X + 1.235168831138153
R-Squared Value: 0.04915071749274069
```

```
The average playback rate of the video
-0.8683614927713552 X + 1.5965247194307803
R-Squared Value: 0.023326348876572145

8.422717751529461 X^2 + -19.902935501387834 X + 12.342611041049725
R-Squared Value: 0.0271132450355398

-446.8687035420573 X^3 + 1530.6127759964359 X^2 + -1747.0355654469408 X + 665.1017637805601
R-Squared Value: 0.046719512277430875

-5006.992594429547 X^4 + 22344.25076670896 X^3 + -37349.357826477135 X^2 + 27713.828278171954 X + -7701.316996005426
R-Squared Value: 0.05070018551786071

92058.37765198684 X^5 + -529141.2099192091 X^4 + 1215430.1843729538 X^3 + -1394605.177406958 X^2 + 799351.1828429047 X + -183096.27245566974
R-Squared Value: 0.05274272660943946
```

```
The number of times the video was rewind-ed
-0.032020750592152966 X + 0.6811988759053531
R-Squared Value: 0.033877726933436825

-0.02071934316445217 X^2 + 0.07586825939794378 X + 0.5718152681096553
R-Squared Value: 0.08211445628421721

-0.007645382834705812 X^3 + 0.04902627648583257 X^2 + -0.09436968646323203 X + 0.6811178762903993
R-Squared Value: 0.09842541746661537

0.0065219714355721 86 X^4 + -0.08820292738482789 X^3 + 0.3701771087794513 X^2 + -0.5722166074523037 X + 0.8966993387120817
R-Squared Value: 0.125587050743 2718

-0.000776391928389487 X^5 + 0.018317628138668954 X^4 + -0.15190334573402622 X^3 + 0.5194744528559562 X^2 + -0.7199390865000003 X + 0.9442142937885102
R-Squared Value: 0.12624813654113298
```

```
The number of times the video was fast-forwarded
-0.004571455131147419 X + 0.627377440011386
R-Squared Value: 0.00045077109180235997

-0.03218659494697687 X^2 + 0.1049832832159027 X + 0.5583486286900956
R-Squared Value: 0.020544865356765807

-0.027137694733087008 X^3 + 0.11290334693963218 X^2 + -0.10826059564885665 X + 0.6383409442527266
R-Squared Value: 0.03025401411445716

-0.025855569311706007 X^4 + 0.15806074786279445 X^3 + -0.3237780478740682 X^2 + 0.277165610612571 X + 0.5375997765638932
R-Squared Value: 0.03844262518714325

-0.03858857014294505 X^5 + 0.3331542103168778 X^4 + -1.0495188109094065 X^3 + 1.454948184769543 X^2 + -0.8228782037156976 X + 0.7537810748763493
R-Squared Value: 0.05268819992433127
```